March 6, 2016
Cameron Palk & Joel Benner
CIS 472 - Machine Learning

Final Project Proposal - What's Cooking from Kaggle

Kaggle recently ended a competition for classifying recipes into the dish's cuisine category based on its ingredients. The data is stored in JSON format with a list of the ingredients as well as the cuisine category label.

Our preprocessing would start by converting the data into a format that we can use in our machine learning algorithms. We are planning on formulating the problem into a 'bag or words' problem so individual ingredients can affect the cuisine classification appropriately.

Another step in our preprocessing will be to combine ingredients which might be spelled a bit differently. Like 'apple' and 'apples' should be viewed as the same ingredient by our algorithms but without this step will be considered different.

We would like to test the results of the following algorithms:

- **Random Forest**
- **Averaged/Vanilla Perceptron Ensemble with Averaging Weights**
- **K-NN (** with the number of clusters based on the number of cuisine categories)

These algorithms will then be tuned with their hyper parameters, we will give explanations behind our choices for these values.

We are interested in learning about Stacked Generalization and Blending. From our understanding, stacked generalization is the idea of taking the predictions of multiple models and using another classifier to combine these predictions with the goal of reducing the generalization error. We aren't sure if we understand stacking well enough to implement it for our problem. We are going to work on the list of algorithms/ensembles listed above first, while also trying to learn about the stacked generalization and blending with the hopes of working it into another ensemble.

We will benchmark our algorithms success based on the accuracy it achieves on the testing data supplied by Kaggle.

The completed Kaggle competition:

kaggle.com/c/whats-cooking